Zefi, Caro, Maxime, November 15th 2019

PRESENTATION

M: You two are currently working together on a specific project at the IISH, right?

C: Actually no, Zefi works on a Sudanese project here at the IISH, and I did my thesis here about social media webarchiving. Also, I am currently constituting a physical archive with the GroenLinks party.

M: When did web archiving appear? Could it be right when social platforms took over?

Z: Well, we've got to distinguish two things here : the digitisation of physical material for its preservation, and the harvest of digital born content. Preservation of digitised material is, of course, a much older thing, while the archiving of born digital materials (material that have not existed in any other forms than digital) is pretty new. If we're talking about web archiving stricly, I'm not sure exactly when and where it started but it's only a few years old.

PROBLEMATICS OF THIS NEW JOBS

M: Because being a rather young discipline, digital archiving must ask you to design protocoles while harvesting, in other words defining your job while doing it? Am I right?

Z: Yes, we've got to do everything. We've got to merge the principles and practices of archiving, and see how they could be applied to our digital research processes.
For some time now, we've been collecting born digital material and social media content. Of course, the question of accessibility and usability of this database rises at some point. Because in the end, that is what it's been harvested for : an archive isn't made to remain untouched in a harddisk for 150years. But so far, nothing is polished enough to reach a satisfying state for public consultation.

M: You mentioned a project about the Sudanese uprising?

Z: We started very recently a project which consists in collecting material from the Sudanese uprising. The events started in the end of the year 2018, and our project started about two months ago. For this, we work with a specialized currator from Sudan, who is himself involved in the movement. He is basically our consultant in the developping process of our collection which is mainly made of digital born materials, things we got from social platforms. We are provided by an Amsterdam based Sudanese radio station.

M: How do you select the content? Do you keep everything you receive? Is automation relevant in your sorting process?

C: We are provided by different people with different status and roles in several social movements. The importance in keeping digital content (even though on the internet and maybe in some cases largely spread out) is that it will not stay there forever. For instance, I researched american social movements, and we noticed that about sixty pourcents of the websites created during the movement were gone. Usually, we can estimate the loss rate of web based digital material of fifty pourcent within two years. Therefore, it's very important for us to harvest all of that in time.

Z: Selection is quite tricky, for multiple reasons, but the main one being the enormous quantity of material to be found online. I've seen from experience that the current approach basically consists in trying to harvest as much as possible. But to simply go through all of a harvest is really time demanding. On Twitter, for instance, the processing of thousands of pictures is necessary but really complicated, if you do it manually. For that, automation can help, sometimes. We are still questioning its relevance. Full or semi automation could be a helpful tool, up to points that depend on the material you harvest, of course. For instance, when harvesting on a website, the most common method is of *crawling,* which is a semi-automated process that allows to acquire specific websites, from specific URLs. But it’s rather hard to get relevant content from that process, as you will easily obtain mistakes, shifted content, etc.

M: Hmm, it seems that it’s a rather different job than what the common imaginary would think of, actually.

C: Yes, it’s very different, very technical. An archivist, before, would focus on paper, which are surely stable documents. In their stack of paper, the archivist would find an order.

M: As I understand your work, you’re cutting slices of an ongoing timeline. I find interesting the necessity to make a still, always, to capture a content that is very much in motion.

Z: Yes, there’s indeed no way to preserve the fluidity of the web. The archived version of it is a different thing, really. It’s not even a copy, *per se*. It’s an accumulation of elements from different moments in time, a frozen slice of digital content, approximating the original. It’s actually a common issue in webarchiving to deal with a sort of temporal incoherence. From the moment you start harvesting a website (a very long process in essence), to these when your harvest is complete, it is often possible that the website has changed. You will then have parts of both old and new website. With social medias, which are often updated, this will often be the case.

C: This makes the job of a web archivist very different than of a classic archivist. A webarchivist will then have to come out with a few choices. What to do with a website that has changed its topic target, for instance?

Z: Although, I still think there’s space for traditional archiving ideals in webarchiving. We are still taught the importance of context, provenance. Without this, you don’t have an archive, but a data dump.

C: Yes. Right now, the main question focuses on the categorization, the accessibility of digital content. When it comes to description, cleaning of content, or in a broader sense, to make it usable, there is no conclusive answer, yet, but there are “accepted ways of doing”. For instance, the Internet Archive (archive.org) has this tool which allows to see websites in timeline, the *wayback machine*. Knowing they can’t archive the whole internet, the Archive still wanted to keep record, starting somewhere. And because they are highly constrained by technicalities, they mostly harvest documents that fell under the public domain.

M: Is that what also happens in at the IISH?

Z: Somehow yes. We cannot harvest all feeds from all social medias, especially of private pages. A few months ago, we coorperated with a group from Egypt. They allowed us to harvest on their facebook page, which was completely private. As we needed access facilities, we had a contract with them. It was very much of a formal procedure, that included privacy policies and all.

M: And if your research leads you to a content that is under copyright, how do you proceed?

Z: Well, at the moment, we are harvesting, but not giving access. And as long as we do not publish, we remain in a safe zone. We’re not violating copyright. But all piece of content is anyway under copyright, its unavoidable. At the moment, we keep this conscious approach, and will not give access to it, because that would basically mean having to clean out all material under rights. Perhaps, the publication of material that was donated asks less questions, since given. But the main problem to me is that we would need to process and render publicly accessible our content now, as research is ongoing. If what we collect is central to the social and political life of the world, what is the use of keeping them closed for decades?

SPACE at the IISH

M: How is a digital archivist affected by the problematic of space?

Z: There’s a general assumption on the fact that digital content would not take up space, this is of course wrong. It is a problem of money, mainly. Storage is of course cheaper than it use to be, but it is still expensive. When you store a few back up copies like we do (we have four of them), you really have to make some wise choices, and because of the scarcity of space, we have to select our content.

M: So technically, how does the IISH spreads out its data?

Z: Our data is not here at the IISH, obviously. Part of it is in a servor in Sciencepark, here in Amsterdam, and the rest is in another location. Currently, we rent out space. That should be enough for a good period of time, but as the IISH is a part of a cluster of institutes, sharing facilities and networking is a long term goal.

FORMAT

M: What happens when harvesting content that hasn't a graphic shape, or user friendly format?

Z: The best form to preserve and give access to the material still is an open question. There are mutliple possibilities, alternatives, that give rise to diverse opportunities. At the IISH, the main audience isn't a general audience. Researchers are our designated community of users. Not all of them are familiar with digital material processes but still it isn't a total discovery for them. At the IISH, we are considering giving at least two formats to all collected data, allowing the maximum users to consult it.

M: And how would that work?

Z: Well, when it comes to social medias for instance, one of our tool allows to run the webpage as it was when it was still online, which makes it easily usable. Another method is to get the raw data of a social media platform, and we then use filters and analyse the content. Both of this methods have *pros* and *cons*. Often the second method appears to be better, for it allows to play around with the data quite a lot, and it doesn't require much storage space, since there's no graphic content along. However, the other approach has the advantage of preserving the context very well : the look and feel is kept, and that could sometimes be a valuable thing to have. Most of us are very used with graphical interfaces to do stuff. But when you work with a digital archive, you cannot resort to such tool, for they simply don't exist. Tools we have are very specifically targeted tasks executors.

M: Do you sometimes come to deal with hybrid or semi digital archives?

C: Yes, if you look at what's been done in the early 2000's, there's still a lot of untouched CDs, floppy disks, etc. There are now several ways figured out when it comes to dealing with semi digital content, but of course not all material has been processed yet. And they aren't processed yet because it isn't urgent or else, but because most often what is to be found inside isn't a streamline. A lot of time is needed to get a coherent use out of that. So instead of rushing, we wait for an opportune moment to take them back out of the shelves. In the end, we're following the technological advancies, always running after the web. As webarchivists, we will always be *achterstaat* (behind).

DIGITAL archivist work NOWADAYS

M: It feels like your senses could be sometimes useless in your work process.

Z: Yes, we're crawling blindfold in the air.

C: It kind of feels like it yes, and not only because we're dealing with digital content, but because we are learning and defining what digital archiving is. What's difficult is that if you look at a paper archive, you can see it, feel it, read it and put it aside. When it comes to digital information, you cannot touch, and sometimes you cannot even see. Which is problematic because an archivist should have a certain idea of the globality of their archive. So not being able to fully relie on your senses limits you.